WORKING WITH DATA

In order to solve the business problem we have to come up with the best suited location to open the medical rehabilitation centre so that it can function with the optimal results. We need the following data to form insights accordingly.

1. District wise count of the total number of quarantined, confirmed, active and recovered cases.
2. District wise count of the total number of people suffering due to mental health like stress, anxiety, depression, sleep impairment and deaddiction cases.
3. District wise division of hotspots and containment zones declared by the government.
4. Latitude and longitude coordinates of those hotspots. This data is required to spot the locations of hotspots on the map, and also to extract venue data from Foursquare API.
5. Venues, particularly related to hospitals so that we can perform clustering near the hotspots.

Data Source and extraction:

All the necessary data regarding the covid-19 pandemic was available on Kerala government official website (http://dashboard.kerala.gov.in/). I prepared a few excel sheets out of the data extracted from the same. Namely "Kerala district wise breakdown" giving the current statistics of the ones who were physically affected," Psychosocial data" giving an insight on the emotionally affected ones and finally the count of LSGs needing special attention which were defined to be the hotspots.

Then the geographical co ordinates of these hotspots were found using the python geocoder package. Using the latitude and longitude co ordinates we can visualize the hotspots on the map to easily find out the badly affected regions.

For the venue data, we use the Foursquare API. Use of foursquare is focused to fetch nearest venue locations so that we can use them to form a cluster. Foursquare API leverages the power of finding nearest venues in a radius and also corresponding coordinates, venue location and names. Since we are particularly interested in the hospital category, we will extract the names of nearby hospitals in order to solve the business problem put forward. We are mainly concentrating on the hospital category because after acute-stage care in hospitals, an intensive program of treatment and rehabilitation is necessary to assist the patient to recover quickly as possible. We are also giving equal importance for psychosocial support needed and planning to start certain state wide volunteer programs on the regions of most concern.

METHODOLOGY

The first and foremost thing to start with is to carefully analyse the data collected because it gives us initial insights and may help to get partial idea of the answers that we are looking to find out from the data.

- We will start with the main data and visualize the order of districts affected in terms of quarantined, active, recovery and death cases. From which we can confirm the regions which has to be noted with at most concern.

- Using the mental health data we will categorize the types of discomfort people are facing. There was an increasing number of suicides reported due to unavailability of alcohol during the lockdown days. Hence this an important matter of concern.

- Using the count of hotspots per district we will use the folium package and explore them on the map. Pythons geocoder package allows us to convert address into geographical coordinates in the form of latitude and longitude. This can provide us the areas to be considered while building the recommender model that can be of maximum benefit. Plotiing them on a map allows us to perform a sanity check to make sure that the geographical coordinates data returned by geocoder are correctly identified.

- With the help of Foursquare API we can get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the hotspots in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each hotspots and examine how many unique categories can be curated from all the returned venues.

- Then, we will analyse by grouping the rows by hotspots and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since, we are analysing the “Hospital” data, we will filter the “Hospital” as venue category.

Finally, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the hotspots into 3 clusters based on their frequency of occurrence for “Hospital”. The results will allow us to identify those regions with least, moderate and highset number of hospitals including the containment zones, that can be incorporated with the insight we drew from the exploratory data analysis of the districts highly in need of a rehab.